THE EFFICIENCY RATINGS OF TEACHERS

C. W. HILL
Superintendent of Schools, Marengo, Illinois

School administrators have long felt the need of a reliable scale for rating the efficiency of teachers. Many have tried, with varying results, scales already devised. Some have become discouraged and given them up as failures, while others have tried modifying schemes with the hope of working out a more satisfactory plan. The writer has attempted to check up the efficiency rating schemes that are in most common use in the elementary schools with the results of the teachers' instruction, in order to discover, if possible, both their weak and strong characteristics. To do this, three steps were taken. First, a preliminary survey was made to determine what sorts of rating schemes are in common use. Second, a comparison was made between the efficiency rating of a teacher and the results of her classroom instruction as shown by standard tests. Third, an analysis was made of the findings to discover as nearly as possible both the agreements and disagreements between the two sets of measurements.

Thirty-four school systems in cities widely distributed over the United States and ranging in population from 10,000 to 800,000 reported. These reports consisted of statements of methods used in determining the efficiency of their elementary teachers, and copies of their printed rules and regulations. The results show that 41 per cent are now using some definite rating scheme. This is an increase over the number found by Boyce in his study made in 1915. In most cases the schemes in use are patterned very largely after either the Boyce or the Engelhardt-Strayer rating cards. Some have made changes in the subheads, but for the most part the main divisions are the same as in the original schemes. Four of the cities reporting (Denver, Colorado; Lincoln, Nebraska; Everett, Washington; and Lewiston, Idaho) use Boyce's Scale in its entirety. A close study of the scales used in the other cities

revealed the fact that the lists of subheads are almost identical. The different administrators disagree only slightly as to the best groupings of these divisions and their weightings.

In some cases a short statement is printed on the rating card explaining just what is to be included under each subhead to guide the person in making his judgment. In 78 per cent of the cases use is made of the letters A, B, C, and D; or E, G, M, and P. Fifty per cent add the fifth grade, E, or V. P. Detroit alone used the plus and minus in connection with the letters. Detroit also was the only one to make use of the human scale method. By this method a teacher of the system is placed to represent each of the different grades of each different quality indicated on the scale, the teacher's efficiency rating being determined by comparison with these representative individuals.

In the main divisions of the rating schemes, *results of instruction* rank first; *classroom management* (including discipline), second; *technique of instruction*, third; *personal equipment* and *co-operation* tied for fourth place; *academic training*, sixth; *professional training*, seventh; *loyalty*, eighth; *experience*, ninth; and *general intelligence*, last. In comparing the results of this survey with those found by Boyce, Ruediger and Strayer, and Moses, it is noted that the displacement of qualities is very slight. In Boyce's work the division, *results of instruction*, is placed first, and in Moses' *technique of instruction* is first (this work does not mention results of instruction). Ruediger and Strayer give second place to *results of instruction* and put *classroom management* first.

To make the comparison between the different efficiency ratings of teachers and the results of their instruction it was necessary to determine a unit that would be comparable in the different systems. The plan devised is based upon three suppositions. First, the results of a teacher's work in three or more classes will be a fair sample of all her work. Second, the results of standard tests are the most reliable measures of the results of instruction to be had at present. And third, the most reliable tests now available are in the fields of arithmetic, penmanship, and spelling. To get a comparable unit it was decided to use the difference between the standards set by the administration for the close of two consecutive

semesters or periods of work. For example, the standard set for sixth-grade arithmetic in a particular school at the end of one semester was 490, and for the end of the next semester, 770. In that case the unit to be covered was 280. The unit was applied as follows: In the school from which the above standards were taken, one class stood at the beginning of a semester at 508 (18 above standard) and at the close of the term at 785, showing a gain of 277, which equals 98 per cent of the unit prescribed, or .98 units. This plan was used with each of the different classes in which the teacher's results were tabulated. For the total score to indicate the results of her instruction, the three different units were averaged. For example, a certain teacher had the following score:

$$
\begin{array}{l}
\text{Penmanship} \dots\dots\dots\dots\dots 100 \text{ per cent or } 1.00 \text{ units} \\
\text{Spelling} \dots\dots\dots\dots\dots\dots\;\, 80 \text{ per cent or } .80 \text{ units} \\
\text{Arithmetic} \dots\dots\dots\dots\dots\;\, 79 \text{ per cent or } .79 \text{ units} \\
\hline
\qquad\qquad 3)259 \text{ per cent or } 2.59 \text{ units} \\
\hline
\qquad\qquad 89\tfrac{1}{3} \text{ per cent or } .89\tfrac{1}{3} \text{ units}
\end{array}
$$

Neither special classes nor those working under exceptionally good or poor conditions were used in making these comparisons.

In gathering these data it was necessary to find schools that would meet the three following conditions. First, they must be willing to co-operate and give all the data they have that would aid the study. Second, they must be schools that are giving their teachers efficiency ratings by people trained for the work; that is, they should have administered the rating schemes a sufficient number of times to insure good technique. Third, both the administration and the teaching force must be giving these ratings sufficient consideration to make them a serious part of their school life. Three systems, Winnetka, Illinois, Gary, Indiana, and Detroit, Michigan, were found to meet all of the above conditions. Data for 135 teachers were secured from these three· places. In gathering the data, blanks were filled out by the administrators of each school, giving information such as presented in Table I.

When the material had been carefully examined for mistakes and omissions, it was tabulated as in Table II. In working out the correlations an efficiency grade A was given a score of 3; a grade

B, score of 2; and grade C, score of 1. The results of the standard tests were transferred just as shown in Table II. The correlations found are shown in Table III. The Pearson formula was used in calculating the correlation between the total numbers; but in all other cases Ayres's short method was used. Two tests were applied to determine whether a random sample had been secured or not. First, the data were divided into halves and the correlation co-efficients determined. The co-efficients for the two halves were found to be .410 and .520 respectively, with a probable error of .046.

TABLE I

RECORD OF RESULTS OF STANDARD TESTS AND TEACHER'S EFFICIENCY GRADE

Teacher Number	Grade	Subject	Date 1st Test	Median of Class	Standard for Class	Date 2d Test	Median of Class	Standard for Class	Efficiency Rating
1......	6B	Arith.	9/19	577	585	1/20	797	781	A
2......	6B	Arith.	9/19	400	585	1/20	653	781	C
7......	6B	Arith.	9/19	770	585	1/20	963	781	B

TABLE II

RETABULATION OF EACH TEACHER'S UNIT SCORES IN EACH OF THE DIFFERENT SUBJECTS FOR THE PURPOSE OF DETERMINING THE FINAL AVERAGE SCORES

Teacher Number	Spelling	Arithmetic	Writing	Average	Efficiency Rating
63..........	.33	.29	1.58	.73	C
69..........	.73	.92	.59	.75	B
78..........	.95	.66	.86	.82	B

Theoretically the correlation would have been between .408 and .500. This indicates that it is highly probable that a random sample had been secured. As a second test the normal probability curve was applied to the results of instruction as shown by the standard tests. The 135 cases were graphed in the form of a frequency polygon which was smoothed the second time. The actual distribution was found to fit very closely to the normal curve. The mode and mean differed only by .0870. The skewness equaled .087. These two tests furnish sufficient evidence to indicate that the data are reliable for the purpose intended.

Three factors at least operate to lower the relationship indicated in Table III. First, standard tests are not perfect, and therefore the results will always contain more or less of error. Second, something besides objective results of instruction such as discipline, co-operation, enthusiasm, etc., go to make up a part of the total of a teacher's efficiency grade. A third is that the administrator usually relies partially upon his subjective judgment in making the ratings.

An examination of the rating schemes used in the three different schools indicates that in Winnetka approximately 75 per cent of the weighting in making the ratings is placed upon the tangible results of instruction; in Gary about $33\frac{1}{3}$ per cent, and in Detroit practically none. This gives almost a perfect correlation between

TABLE III

THE VARIOUS CORRELATION CO-EFFICIENTS DETERMINED

Data compared	Correlations
1. Test scores and ratings, all schools (135 cases)	.454
2. Test scores and ratings, Winnetka	.450
3. Test scores and ratings, Gary	.240
4. Test scores and ratings, Detroit	.190
5. Arithmetic scores and ratings, Gary	.630
6. Spelling scores and ratings, Gary	.060
7. Arithmetic scores and ratings, Detroit	.080
8. Spelling scores and ratings, Detroit	.190

the correlations shown in Table III and the percentage of weight given to the results of instruction. The administrator's reports regarding the attitude of the teachers who were given efficiency ratings were as follows: Detroit, "Several were dissatisfied." Gary, "From 2 per cent to 5 per cent asked for conferences with the administrators." Winnetka, "No complaints."

A study of the rating schemes used brings out the fact that different administrators emphasize different things in determining a teacher's efficiency rating. In Winnetka it is ability to produce tangible results of instruction. Detroit takes almost the opposite view, placing the tangible results last and emphasizing executive ability, leadership, personality, etc. Gary, where the practice comes more nearly being typical, takes a middle ground and gives the results of instruction about one-third of the weighting, distribut-

ing the remaining two-thirds among co-operation, loyalty, leadership, etc. It is evident that the same scheme cannot be used with these varying methods and get the results expected by the administrators.

Any rating scale needs to be administered with much caution. At best such a device will only serve to indicate what may be true. Further investigation should be made before any serious decisions are made. In the first place, the teacher should know in advance by what means and upon what grounds she is to be judged. The person making the rating should take into consideration the conditions under which the instructor has had to work. The kind of pupils being taught, physical conditions, etc., all affect the results to be achieved.

The answers given above to the question about teachers' dissatisfaction with their ratings show that the complaints decrease as more weight is placed upon the tangible results of instruction. At least three factors aid in producing satisfaction, or the lack of it, among the teachers rated. First, the teacher might have come more nearly to an agreement with the administration in regard to the elements that should receive consideration in making up the ratings. Second, one administrator might be more skilful in managing his teachers than some others, or more accurate in his judgments. And third, the ratings in one case might be based to a greater extent upon objective evidence, whereas in another case subjective judgments would get more consideration.

This survey, when compared with previous studies, indicates that an increasing number of school administrators are making use of some definite rating plan. It shows that with the operation of efficiency-rating plans, the dissatisfaction among teachers decreases with the increase in the use of objective data. It also shows that the trend is to place more and more emphasis upon the tangible results of instruction in rating teachers.